Argonne NATIONAL LABORATORY

… for a brighter future

Visual Characterization of I/O System Behavior for High-End Computing

Kwan-Liu Ma (PI), Chris Muelder (UC Davis)

Kamil Iskra, Pete Beckman (CI UChicago/ANL)

U.S. Department of Energy

UChicago ▶ Argonne LLC

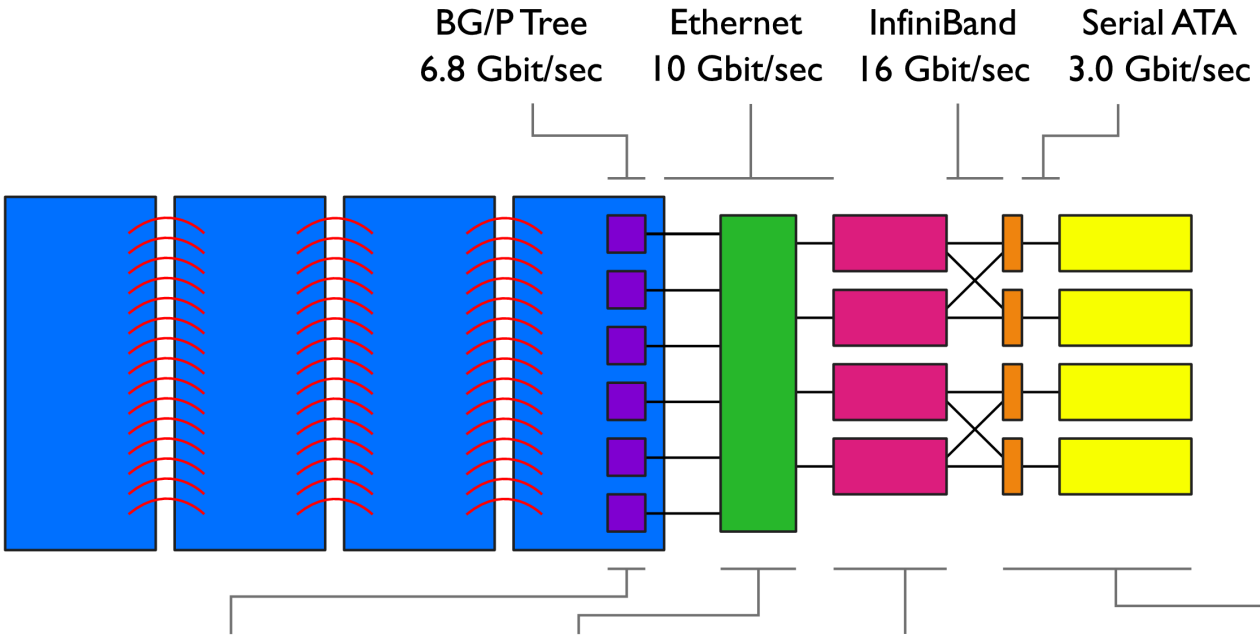
Office of Science U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

Contents

- Motivation
- Plan
- Data Gathering
- Visualization
- Conclusion

HEC Complexity: Hardware



Gateway nodes run parallel file system client software and forward I/O operations from HPC clients.

640 Quad core PowerPC 450 nodes with 2 Gbytes of RAM each

Commodity network primarily carries storage traffic.

900+ port 10 Gigabit Ethernet Myricom switch complex

Storage nodes run parallel file system software and manage incoming FS traffic from gateway nodes.

136 two dual core Opteron servers with 8 Gbytes of RAM each

Enterprise storage controllers and large racks of disks are connected via InfiniBand or Fibre Channel.

17 DataDirect S2A9900 controller pairs with 480 1 Tbyte drives and 8 InfiniBand ports per pair

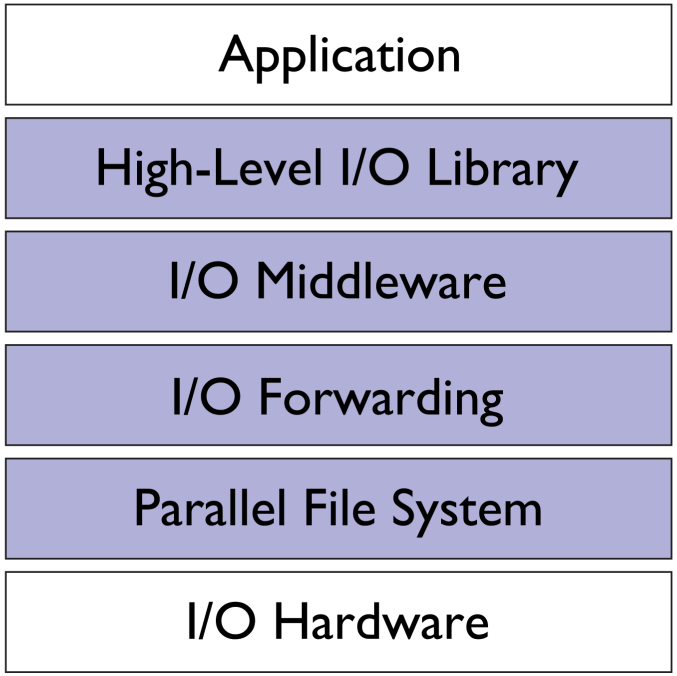
HEC Complexity: Software

High-Level I/O Library
maps application abstractions onto storage abstractions and provides data portability.

HDF5, Parallel netCDF, ADIOS

I/O Forwarding
bridges between app. tasks and storage system and provides aggregation for uncoordinated I/O.

IBM ciod



I/O Middleware
organizes accesses from many processes, especially those using collective I/O.

MPI-IO

Parallel File System
maintains logical space and provides efficient access to data.

PVFS, PanFS, GPFS, Lustre

Understanding Performance Behavior of HEC Systems

- Lots of research on application performance analysis and debugging tools:
  - Jumpshot, ParaGraph, Vampir, TAU
- The needs of system software developers often overlooked:
  - impossible to effectively tune the OS without
  - high degress of efficiency unattainable by applications

Plan

We will develop software infrastructure to provide end-to-end analysis and visualization of I/O system software.

Develop/improve/deploy:

- end-to-end, scalable tracing integrated into the I/O system (MPI-IO, I/O forwarding, file systems),
- new visual representations and analysis techniques for inspecting traces and extracting knowledge, scalable to very large systems and integrable with existing techniques.

Data Gathering

Capture data:

- at HEC scales,
- from all layers of the I/O software stack,
- correlated to allow associating operations between layers,
- including information on failures and performance anomalies,
- using a general model applicable across a range of systems.

Large Scale Tracing

Adopt successful techniques from the field rather than build new ones, focus on gaps in existing tools:

- HPCT-IO, IOT,
- TAU,
- Sandia's lightweight tracing in SYSIO,
- MPE.

Initial aim:

- gather logs of full-scale application runs performed on Intrepid (Argonne's 557 TF BG/P)

Comprehensive Data Gathering

Cover all layers of the I/O software stack:

- a high-level I/O library,
- POSIX and MPI-IO (PMPI),
- I/O forwarding,
- filesystem clients and servers,
- hardware?

End-to-End Capture of Behavior

Correlate data obtained from individual components at different layers into a coherent whole.

- augment all components from the previous slide to enable the correlation between layers,
- API/protocol extensions required.

We want to be able to match operations on I/O servers back to the application I/O call that initiated them.

Visualization

Create the information visualization algorithms and applications necessary to provide insight into I/O system behavior under a wide variety of workloads.

- interactive drill-down techniques,
- visual space design for multidimensional linking,
- metric design for data filtering and abstraction,
- application-driven strategies.

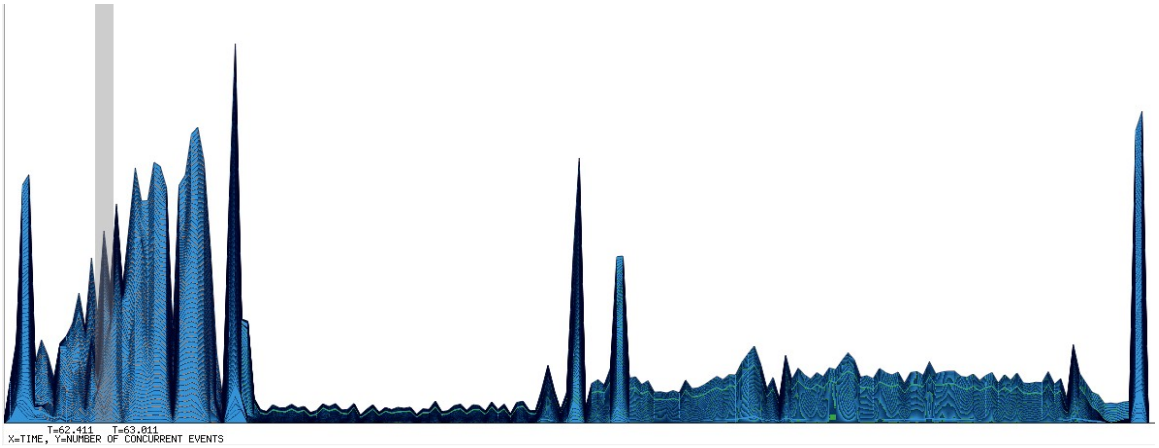
Interactive Drill-Down Techniques

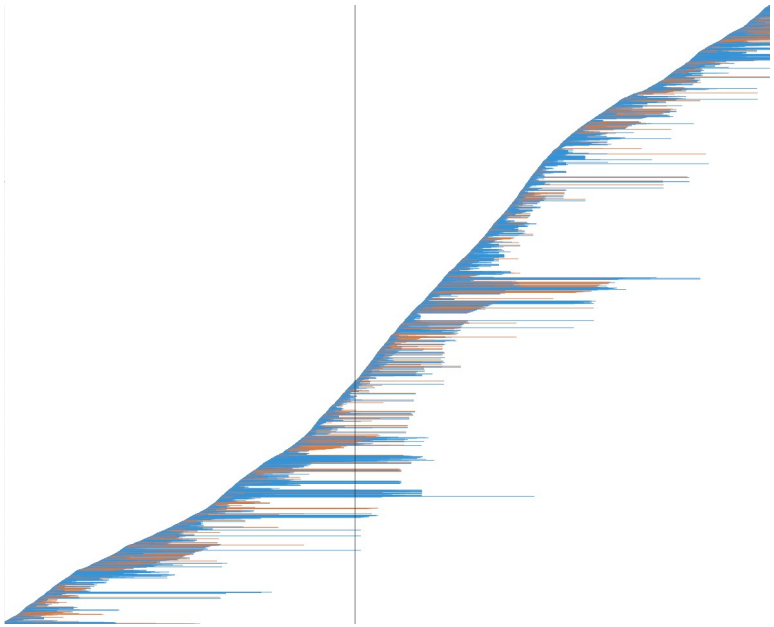
Given the scale, any visual analysis tool needs to provide summary information and the ability to see detailed data in context.

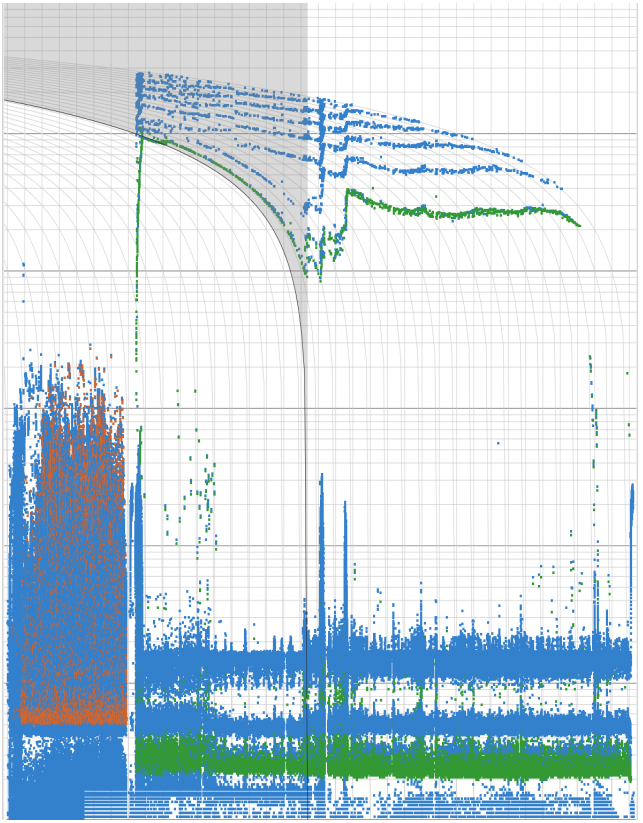
- multi-scale visualization approach,
- intuitive visualization interface,
- study I/O strategies in a high-dimensional optimization space.

Visualize various information aggregations at different levels of detail in a multidimensional space by choosing from a collection of visual transformations.

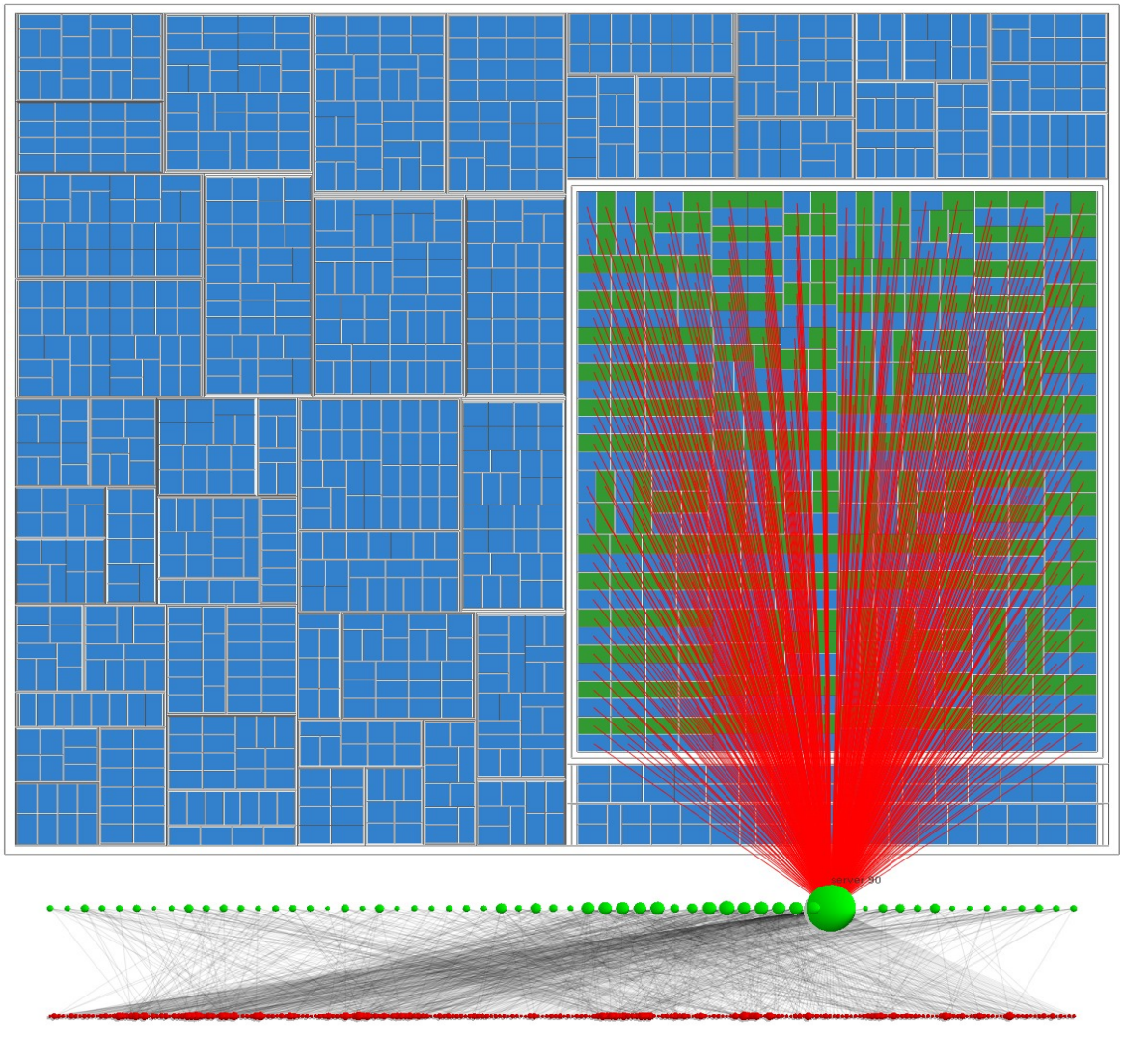
Preliminary Work







Preliminary Work



Visual Space Design

Provide insight into correlating patterns by presenting information from multiple levels of I/O communication in a coherent and linked manner:

- communication paths and system topology,
- bottlenecks in intermediate layers,
- file access patterns on I/O servers,
- tracking the flow of data.

Metric Design

Even on relatively small systems log data is in XX GB range.

We will seek to mine collected data that is too large before visualization.

- standalone metrics measured on individual server/process/connection (load, bandwidth, etc),
- graph theory based (bipartite degree centrality, etc),
- clustering based (two-mode clustering, distance based clustering).

Conclusion

This project has not started yet.

Expected outcomes:

- provide more thorough understanding of unexpected storage system behavior (bottlenecks, etc),
- enable more effective performance debugging through the pinpointing of specific underperforming storage components,
- improve the understanding of the mapping of application data structures to the storage ones, and its impact.

(this slide intentionally left blank)